(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 118350463 B
(45)授权公告日 2024.09.27

(21)申请号 202410779372.2
(22)申请日 2024.06.17
(65)同一申请的已公布的文献号
申请公布号 CN 118350463 A
(43)申请公布日 2024.07.16
(73)专利权人 恒生电子股份有限公司
地址 310056 浙江省杭州市滨江区滨兴路1888号43层
(72)发明人 陈奕名 刘海燕 林金曙
(74)专利代理机构 北京智信禾专利代理有限公司 11637
专利代理师 张瑞
(51)Int.Cl.
G06N 5/025(2023.01)
G06F 40/35(2020.01)

(56)对比文件
CN 117786070 A,2024.03.29

审查员 胡振洲

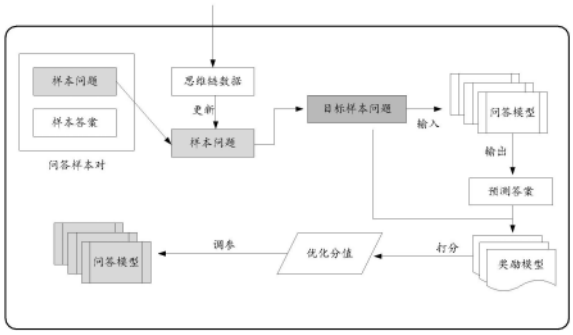
权利要求书3页 说明书17页 附图6页

(54)发明名称

问答模型训练方法、文本处理方法及奖励模型训练方法

(57)摘要

本说明书实施例提供问答模型训练方法、文本处理方法及奖励模型训练方法,其中所述问答模型训练方法包括:在问答样本对中提取样本问题,并确定所述样本问题对应的思维链数据;利用所述思维链数据将所述样本问题更新为目标样本问题,并将所述目标样本问题输入至初始问答模型进行处理,获得预测答案;利用所述初始问答模型关联的奖励模型根据所述问答样本对中的样本答案,对所述预测答案进行打分,获得优化分值;基于所述优化分值对所述初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型。



CN 118350463 B

1.一种问答模型训练方法,其特征在于,包括:

在问答样本对中提取样本问题,并确定所述样本问题对应的思维链数据;

在所述样本问题中确定与所述思维链数据匹配的待变更字单元,利用所述思维链数据对所述待变更字单元进行更新,获得目标样本问题,并将所述目标样本问题输入至初始问答模型进行处理,获得预测答案,其中,所述预测答案包含所述样本问题对应的预测答案文本以及所述样本问题对应的解答思路文本;

利用所述初始问答模型关联的奖励模型根据所述问答样本对中的样本答案,对所述预测答案进行打分,获得优化分值;

基于所述优化分值对所述初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型;

其中,所述奖励模型的训练,包括:获取奖励样本以及所述奖励样本对应的分值向量样本序列,其中,所述分值向量样本序列包含所述奖励样本中每个字单元对应的标准分值向量;将所述奖励样本输入至初始奖励模型进行打分,获得分值向量预测序列,其中,所述分值向量预测序列包含所述奖励样本中每个字单元对应的预测分值向量;利用预设的目标损失函数对所述分值向量样本序列和所述分值向量预测序列计算损失值,并基于所述损失值对所述初始奖励模型进行调参,直至获得满足奖励训练停止条件的奖励模型;其中,所述目标损失函数包含对最后一个字单元对应的目标分值向量进行约束的向量约束项。

2.根据权利要求1所述的问答模型训练方法,其特征在于,所述确定所述样本问题对应的思维链数据,包括:

确定所述样本问题对应的问题类型,在目标数据库中选择所述问题类型匹配的候选思维链数据,作为思维链数据。

3.根据权利要求1所述的问答模型训练方法,其特征在于,所述利用所述初始问答模型关联的奖励模型根据所述问答样本对中的样本答案,对所述预测答案进行打分,获得优化分值,包括:

加载所述初始问答模型关联的奖励模型,以及将所述目标样本问题和所述预测答案拼接为优化文本;

在所述问答样本对中提取样本答案,并利用所述奖励模型根据所述样本答案对所述优化文本进行打分,获得优化分值。

4.根据权利要求1所述的问答模型训练方法,其特征在于,所述基于所述优化分值对所述初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型,包括:

根据所述优化分值和所述样本答案针对所述初始问答模型构建模型优化策略;

按照所述模型优化策略对所述初始问答模型进行调参,并检测调参后的初始问答模型是否满足训练停止条件;

若否,将调参后的初始问答模型作为初始问答模型,并执行在问答样本对中提取样本问题的步骤;

若是,将调参后的初始问答模型作为目标问答模型。

5.根据权利要求1所述的问答模型训练方法,其特征在于,还包括:

确定所述初始奖励模型预设的初始损失函数;

响应于针对所述初始奖励模型提交的模型优化请求对所述初始损失函数进行更新,其

中,所述模型优化请求用于在所述初始损失函数中添加语义学习信息;

根据更新结果确定目标损失函数,并执行利用预设的目标损失函数对所述分值向量样本序列和所述分值向量预测序列计算损失值的步骤。

6.一种文本处理方法,其特征在于,包括:

接收客户端上传的问题文本;

确定所述问题文本对应的问题领域信息,并选择所述问题领域信息匹配的目标问答模型,其中,所述目标问答模型通过权利要求1至5任一项所述的问答模型训练方法获得;

将所述问题文本输入至所述目标问答模型进行处理,获得答案文本,并将所述答案文本反馈至所述客户端。

7.一种奖励模型训练方法,其特征在于,包括:

获取奖励样本以及所述奖励样本对应的分值向量样本序列,其中,所述分值向量样本序列包含所述奖励样本中每个字单元对应的标准分值向量;

将所述奖励样本输入至初始奖励模型进行打分,获得分值向量预测序列,其中,所述分值向量预测序列包含所述奖励样本中每个字单元对应的预测分值向量;

确定所述初始奖励模型预设的目标损失函数,其中,所述目标损失函数包含对最后一个字单元对应的目标分值向量进行约束的向量约束项;

利用所述目标损失函数对所述分值向量样本序列和所述分值向量预测序列进行计算,根据计算结果对所述初始奖励模型进行调参,直至获得满足奖励训练停止条件的奖励模型。

8.一种问答模型训练装置,其特征在于,包括:

提取模块,被配置为在问答样本对中提取样本问题,并确定所述样本问题对应的思维链数据;

更新模块,被配置为在所述样本问题中确定与所述思维链数据匹配的待变更字单元,利用所述思维链数据对所述待变更字单元进行更新,获得目标样本问题,并将所述目标样本问题输入至初始问答模型进行处理,获得预测答案,其中,所述预测答案包含所述样本问题对应的预测答案文本以及所述样本问题对应的解答思路文本;

打分模块,被配置为利用所述初始问答模型关联的奖励模型根据所述问答样本对中的样本答案,对所述预测答案进行打分,获得优化分值;

调参模块,被配置为基于所述优化分值对所述初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型;

其中,所述奖励模型的训练,包括:获取奖励样本以及所述奖励样本对应的分值向量样本序列,其中,所述分值向量样本序列包含所述奖励样本中每个字单元对应的标准分值向量;将所述奖励样本输入至初始奖励模型进行打分,获得分值向量预测序列,其中,所述分值向量预测序列包含所述奖励样本中每个字单元对应的预测分值向量;利用预设的目标损失函数对所述分值向量样本序列和所述分值向量预测序列计算损失值,并基于所述损失值对所述初始奖励模型进行调参,直至获得满足奖励训练停止条件的奖励模型;其中,所述目标损失函数包含对最后一个字单元对应的目标分值向量进行约束的向量约束项。

9.一种文本处理装置,其特征在于,包括:

接收模块,被配置为接收客户端上传的问题文本;

确定模块,被配置为确定所述问题文本对应的问题领域信息,并选择所述问题领域信息匹配的目标问答模型,其中,所述目标问答模型通过权利要求1至5任一项所述的问答模型训练方法获得;

发送模块,被配置为将所述问题文本输入至所述目标问答模型进行处理,获得答案文本,并将所述答案文本反馈至所述客户端。

10.一种奖励模型训练装置,其特征在于,包括:

获取样本模块,被配置为获取奖励样本以及所述奖励样本对应的分值向量样本序列,其中,所述分值向量样本序列包含所述奖励样本中每个字单元对应的标准分值向量;

打分样本模块,被配置为将所述奖励样本输入至初始奖励模型进行打分,获得分值向量预测序列,其中,所述分值向量预测序列包含所述奖励样本中每个字单元对应的预测分值向量;

确定函数模块,被配置为确定所述初始奖励模型预设的目标损失函数,其中,所述目标损失函数包含对最后一个字单元对应的目标分值向量进行约束的向量约束项;

调参模型模块,被配置为利用所述目标损失函数对所述分值向量样本序列和所述分值向量预测序列进行计算,根据计算结果对所述初始奖励模型进行调参,直至获得满足奖励训练停止条件的奖励模型。

11.一种计算设备,其特征在于,包括:

存储器和处理器;

所述存储器用于存储计算机可执行指令,所述处理器用于执行所述计算机可执行指令,该计算机可执行指令被处理器执行时实现权利要求1至7任意一项所述方法的步骤。

12.一种计算机可读存储介质,其特征在于,其存储有计算机可执行指令,该计算机可执行指令被处理器执行时实现权利要求1至7任意一项所述方法的步骤。

13.一种计算机程序产品,其特征在于,包括计算机程序或指令,该计算机程序或指令被处理器执行时实现权利要求1至7任意一项所述方法的步骤。

问答模型训练方法、文本处理方法及奖励模型训练方法

技术领域

[0001] 本说明书实施例涉及机器学习技术领域,特别涉及问答模型训练方法、文本处理方法及奖励模型训练方法。

背景技术

[0002] 随着计算机技术的发展,大模型在越来越多的场景中得以应用。该类模型通过使用大量的数据和计算资源,能够在各种任务上达到令用户满足的效果。如文本生成、文本分类、命名实体识别、情感分析等,都可以通过训练好的大模型实现。现有技术中,大模型的训练过程中,强化学习是比较关键的部分;强化学习作为一种机器学习方法,其可以通过使模型与环境交互,根据环境给出的奖励来学习和优化模型。然而,强化学习过程中的瓶颈在于奖励模型,奖励模型作为强化学习的关键,其决定了模型在环境中采取行动后所获取的奖励,如果奖励模型能够精准反映出行动好坏程度,则强化学习即可有效的优化大模型。但是,由于实际应用环境比较复杂,奖励模型很难精准反映出当前环境的好坏,致使在此基础上优化出的大模型也不具有更好的预测性能,因此亟需一种有效的方案以解决上述问题。

发明内容

[0003] 有鉴于此,本说明书实施例提供了一种问答模型训练方法。本说明书一个或者多个实施例同时涉及一种文本处理方法,一种奖励模型训练方法,一种问答模型训练装置,一种文本处理装置,一种奖励模型训练装置,一种计算设备,一种计算机可读存储介质以及一种计算机程序产品,以解决现有技术中存在的技术缺陷。

[0004] 根据本说明书实施例的第一方面,提供了一种问答模型训练方法,包括:

[0005] 在问答样本对中提取样本问题,并确定所述样本问题对应的思维链数据;

[0006] 利用所述思维链数据将所述样本问题更新为目标样本问题,并将所述目标样本问题输入至初始问答模型进行处理,获得预测答案;

[0007] 利用所述初始问答模型关联的奖励模型根据所述问答样本对中的样本答案,对所述预测答案进行打分,获得优化分值;

[0008] 基于所述优化分值对所述初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型。

[0009] 根据本说明书实施例的第二方面,提供了一种文本处理方法,包括:

[0010] 接收客户端上传的问题文本;

[0011] 确定所述问题文本对应的问题领域信息,并选择所述问题领域信息匹配的目标问答模型,其中,所述目标问答模型通过所述问答模型训练方法获得;

[0012] 将所述问题文本输入至所述目标问答模型进行处理,获得答案文本,并将所述答案文本反馈至所述客户端。

[0013] 根据本说明书实施例的第三方面,提供了一种奖励模型训练方法,包括:

[0014] 获取奖励样本以及所述奖励样本对应的分值向量样本序列,其中,所述分值向量

样本序列包含所述奖励样本中每个字单元对应的标准分值向量；

[0015] 将所述奖励样本输入至初始奖励模型进行打分，获得分值向量预测序列，其中，所述分值向量预测序列包含所述奖励样本中每个字单元对应的预测分值向量；

[0016] 确定所述初始奖励模型预设的目标损失函数，其中，所述目标损失函数包含对目标分值向量进行约束的向量约束项；

[0017] 利用所述目标损失函数对所述分值向量样本序列和所述分值向量预测序列进行计算，根据计算结果对所述初始奖励模型进行调参，直至获得满足奖励训练停止条件的奖励模型。

[0018] 根据本说明书实施例的第四方面，提供了一种问答模型训练装置，包括：

[0019] 提取模块，被配置为在问答样本对中提取样本问题，并确定所述样本问题对应的思维链数据；

[0020] 更新模块，被配置为利用所述思维链数据将所述样本问题更新为目标样本问题，并将所述目标样本问题输入至初始问答模型进行处理，获得预测答案；

[0021] 打分模块，被配置为利用所述初始问答模型关联的奖励模型根据所述问答样本对中的样本答案，对所述预测答案进行打分，获得优化分值；

[0022] 调参模块，被配置为基于所述优化分值对所述初始问答模型进行调参，直至获得满足训练停止条件的目标问答模型。

[0023] 根据本说明书实施例的第五方面，提供了一种文本处理装置，包括：

[0024] 接收模块，被配置为接收客户端上传的问题文本；

[0025] 确定模块，被配置为确定所述问题文本对应的问题领域信息，并选择所述问题领域信息匹配的目标问答模型，其中，所述目标问答模型通过所述问答模型训练方法获得；

[0026] 发送模块，被配置为将所述问题文本输入至所述目标问答模型进行处理，获得答案文本，并将所述答案文本反馈至所述客户端。

[0027] 根据本说明书实施例的第六方面，提供了一种奖励模型训练装置，包括：

[0028] 获取样本模块，被配置为获取奖励样本以及所述奖励样本对应的分值向量样本序列，其中，所述分值向量样本序列包含所述奖励样本中每个字单元对应的标准分值向量；

[0029] 打分样本模块，被配置为将所述奖励样本输入至初始奖励模型进行打分，获得分值向量预测序列，其中，所述分值向量预测序列包含所述奖励样本中每个字单元对应的预测分值向量；

[0030] 确定函数模块，被配置为确定所述初始奖励模型预设的目标损失函数，其中，所述目标损失函数包含对目标分值向量进行约束的向量约束项；

[0031] 调参模型模块，被配置为利用所述目标损失函数对所述分值向量样本序列和所述分值向量预测序列进行计算，根据计算结果对所述初始奖励模型进行调参，直至获得满足奖励训练停止条件的奖励模型。

[0032] 根据本说明书实施例的第七方面，提供了一种计算设备，包括：

[0033] 存储器和处理器；

[0034] 所述存储器用于存储计算机可执行指令，所述处理器用于执行所述计算机可执行指令，该计算机可执行指令被处理器执行时实现上述问答模型训练方法、文本处理方法或奖励模型训练方法的步骤。

[0035] 根据本说明书实施例的第八方面,提供了一种计算机可读存储介质,其存储有计算机可执行指令,该指令被处理器执行时实现上述问答模型训练方法、文本处理方法或奖励模型训练方法的步骤。

[0036] 根据本说明书实施例的第九方面,提供了一种计算机程序产品,包括计算机程序或指令,该计算机程序或指令被处理器执行时实现上述问答模型训练方法、文本处理方法或奖励模型训练方法的步骤。

[0037] 本实施例提供的问答模型训练方法,为了能够在问答模型训练过程中,可以通过奖励模型精准反映当前样本对训练问答模型后问答模型的预测精度,可以先在问答样本对中提取样本问题,并确定样本问题对应的思维链数据;在此基础上,为了能够使得简短问答依旧可以强化训练问答模型,此时可以利用思维链数据将样本问题更新为目标样本问题,实现通过目标样本问题体现更多的文本内容,此后即可将目标样本问题输入至初始问答模型进行处理,获得预测答案;在此基础上,由于样本问题被更新为目标样本问题,因此奖励模型可以在进行模型预测精度打分时,可以给出能够精准反映模型预测精度的分值,即利用初始问答模型关联的奖励模型根据问答样本对中的样本答案,对预测答案进行打分,即可获得优化分值;最后再基于优化分值对初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型即可。实现在问答模型的强化学习阶段,可以通过思维链数据更新样本问题,使得样本问题携带更丰富的文本信息,以此经过问答模型预测后,可以使得奖励模型针对样本问题给出更加精准的打分,从而通过打分结果精准反映出问答模型需要优化的程度,以此为基础进行后续训练,可以有效的提高模型训练精度,以实现在应用阶段,可以满足下游业务使用。

附图说明

[0038] 图1是本说明书一个实施例提供的一种问答模型训练方法的示意图;

[0039] 图2是本说明书一个实施例提供的一种问答模型训练方法的流程图;

[0040] 图3是本说明书一个实施例提供的一种问答模型训练方法中奖励模型打分结果的示意图;

[0041] 图4是本说明书一个实施例提供的一种文本处理方法的流程图;

[0042] 图5是本说明书一个实施例提供的一种奖励模型训练方法的流程图;

[0043] 图6是本说明书一个实施例提供的一种问答模型训练方法的处理过程流程图;

[0044] 图7是本说明书一个实施例提供的一种问答模型训练装置的结构示意图;

[0045] 图8是本说明书一个实施例提供的一种文本处理装置的结构示意图;

[0046] 图9是本说明书一个实施例提供的一种奖励模型训练装置的结构示意图;

[0047] 图10是本说明书一个实施例提供的一种计算设备的结构框图。

具体实施方式

[0048] 在下面的描述中阐述了很多具体细节以便于充分理解本说明书。但是本说明书能够以很多不同于在此描述的其它方式来实施,本领域技术人员可以在不违背本说明书内涵的情况下做类似推广,因此本说明书不受下面公开的具体实施的限制。

[0049] 在本说明书一个或多个实施例中使用的术语是仅仅出于描述特定实施例的目的,

而非旨在限制本说明书一个或多个实施例。在本说明书一个或多个实施例和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式,除非上下文清楚地表示其他含义。还应当理解,本说明书一个或多个实施例中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[0050] 应当理解,尽管在本说明书一个或多个实施例中可能采用术语第一、第二等来描述各种信息,但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如,在不脱离本说明书一个或多个实施例范围的情况下,第一也可以被称为第二,类似地,第二也可以被称为第一。取决于语境,如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

[0051] 此外,需要说明的是,本说明书一个或多个实施例所涉及的用户信息(包括但不限于用户设备信息、用户个人信息等)和数据(包括但不限于用于分析的数据、存储的数据、展示的数据等),均为经用户授权或者经过各方充分授权的信息和数据,并且相关数据的收集、使用和处理需要遵守相关国家和地区的相关法律法规和标准,并提供有相应的操作入口,供用户选择授权或者拒绝。

[0052] 本说明书一个或多个实施例中,大模型是指具有大规模模型参数的深度学习模型,通常包含上亿、上百亿、上千亿、上万亿甚至十万亿以上的模型参数。大模型又可以称为基石模型/基础模型(Foundation Model),通过大规模无标注的语料进行大模型的预训练,产出亿级以上参数的预训练模型,这种模型能适应广泛的下游任务,模型具有较好的泛化能力,例如大规模语言模型(Large Language Model, LLM)、多模态预训练模型(multi-modal pre-training model)等。

[0053] 大模型在实际应用时,仅需少量样本对预训练模型进行微调即可应用于不同的任务中,大模型可以广泛应用于自然语言处理(Natural Language Processing,简称NLP)、计算机视觉等领域,具体可以应用于如视觉问答(Visual Question Answering,简称VQA)、图像描述(Image Caption,简称IC)、图像生成等计算机视觉领域任务,以及基于文本的情感分类、文本摘要生成、机器翻译等自然语言处理领域任务,大模型主要的应用场景包括数字助理、智能机器人、搜索、在线教育、办公软件、电子商务、智能设计等。

[0054] 在本说明书中,提供了一种问答模型训练方法。本说明书一个或者多个实施例同时涉及一种文本处理方法,一种奖励模型训练方法,一种问答模型训练装置,一种文本处理装置,一种奖励模型训练装置,一种计算设备,一种计算机可读存储介质以及一种计算机程序产品,在下面的实施例中逐一进行详细说明。

[0055] 参见图1所示的示意图,本实施例提供的问答模型训练方法,为了能够在问答模型训练过程中,可以通过奖励模型精准反映当前样本对训练问答模型后问答模型的预测精度,可以先在问答样本对中提取样本问题,并确定样本问题对应的思维链数据;在此基础上,为了能够使得简短问答依旧可以强化训练问答模型,此时可以利用思维链数据将样本问题更新为目标样本问题,实现通过目标样本问题体现更多的文本内容,此后即可将目标样本问题输入至初始问答模型进行处理,获得预测答案;在此基础上,由于样本问题被更新为目标样本问题,因此奖励模型可以在进行模型预测精度打分时,可以给出能够精准反映模型预测精度的分值,即利用初始问答模型关联的奖励模型根据问答样本对中的样本答案,对预测答案进行打分,即可获得优化分值;最后再基于优化分值对初始问答模型进行调

参,直至获得满足训练停止条件的目标问答模型即可。实现在问答模型的强化学习阶段,可以通过思维链数据更新样本问题,使得样本问题携带更丰富的文本信息,以此经过问答模型预测后,可以使得奖励模型针对样本问题给出更加精准的打分,从而通过打分结果精准反映出问答模型需要优化的程度,以此为基础进行后续训练,可以有效的提高模型训练精度,以实现在应用阶段,可以满足下游业务使用。

[0056] 参见图2,图2示出了根据本说明书一个实施例提供的一种问答模型训练方法的流程图,具体包括以下步骤。

[0057] 步骤S202,在问答样本对中提取样本问题,并确定所述样本问题对应的思维链数据。

[0058] 本实施例提供的问答模型训练方法,可以应用于任意问答场景应用的大模型,比如金融场景中针对金融问题进行答复的大模型;再比如,教学场景中针对不同科目问题进行答复的大模型;再比如,开发场景中针对开发问题进行答复的大模型等。本实施例在此不做任何限定。

[0059] 具体的,问答样本对具体是指针对目标领域关联的问答模型进行训练所使用的样本对,其有样本问题和样本答案组成,样本问题作为模型输入,样本答案作为标签,用于在模型训练阶段或者强化学习阶段使用。实际应用中,样本问题和样本对以文本形式存在并训练问答模型。相应的,思维链数据具体是指可以对样本问题进行改装的数据,其目的用于对简短样本问题进行丰富化处理,使得样本问题可以具有更丰富的文本信息,从而便于后续进行模型训练使用,使得模型更容易捕捉文本内的语义关联关系。也就是说,思维链数据可以理解为对样本问题进行更新使用的数据结构,其用于针对样本问题进行替换、更新或者增加字符的方式,使得样本问题可以变更为文本信息更为丰富的目标样本问题,以便于后续使用,也可以理解为思维链数据是在样本问题中增加使模型可以进行思考的问题内容,从容使得模型可以学习该知识,提高预测精度和泛化能力。可以理解为使简短样本问题复杂化,促使模型捕捉复杂化文本中的语义关联关系,以提高模型训练精度。

[0060] 基于此,为了能够在问答模型训练过程中,可以通过奖励模型精准反映当前样本对训练问答模型后问答模型的预测精度,可以先在问答样本对中提取样本问题,并确定样本问题对应的思维链数据;在此基础上,为了能够使得简短问答依旧可以强化训练问答模型,此时可以利用思维链数据将样本问题更新为目标样本问题,实现通过目标样本问题体现更多的文本内容,此后即可将目标样本问题输入至初始问答模型进行处理,获得预测答案;在此基础上,由于样本问题被更新为目标样本问题,因此奖励模型可以在进行模型预测精度打分时,可以给出能够精准反映模型预测精度的分值,即利用初始问答模型关联的奖励模型根据问答样本对中的样本答案,对预测答案进行打分,即可获得优化分值;最后再基于优化分值对初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型即可。

[0061] 进一步的,在确定样本问题对应的思维链数据时,为了能够保证更新后的样本问题可以用于模型训练,并且避免引入冗余信息影响模型训练精度,可以按照问题类型选择思维链数据使用。本实施例中,具体实现方式如下:

[0062] 确定所述样本问题对应的问题类型,在目标数据库中选择所述问题类型匹配的候选思维链数据,作为思维链数据;

[0063] 具体的,问题类型具体是指定位问题所属领域的类型,不同的类型对应不同的思

维链数据。相应的,目标数据库具体是指存储多种问题类型关联候选思维链数据的数据库。相应的,候选思维链数据具体是指在目标数据库中针对样本问题选择到的思维链数据。具体实施时,选择候选思维链数据的操作可以采用计算文本相似度方式实现,即通过计算样本问题与思维链数据之间的文本相似度,之后按照相似度大小进行排序,选择优先级最高的候选思维链数据作为样本问题对应的思维链数据即可。

[0064] 基于此,在从问答样本对中筛选到样本问题后,可以先确定样本问题对应的问题类型,此后可以在存储大量思维链数据库的目标数据库中,选择问题类型匹配的候选思维链数据,作为思维链数据;以便后续将样本问题更新为目标样本问题,从而实现对问答模型的训练。

[0065] 举例说明,获取训练问答模型的问答样本对,其中包含样本问题{1*1+2*2的结果是多少?只回答结果}以及样本答案{5};在此基础上,为了提高问答模型的预测精度,此时可以加载匹配样本问题的思维链数据,如{需要回答思考过程},以便后续可以结合思维链数据对样本问题进行变更,使得样本问题可以包含更为丰富的问题信息,促使在问答模型训练阶段,可以使得奖励模型结合问题和答案给出精准评分,从而训练出满足使用需求的问答模型使用。

[0066] 实际应用中,针对不同的场景可以设置不同的思维链数据使用,比如在金融场景下,可以在问题末尾统一加“返回思考过程,并按照要求格式输出结果”思维链数据,使得模型可以结合新的问题给出新的答案。此过程中,可以通过训练好的大模型给出一组答案,以及通过需要训练的大模型给出另一组答案,此后,通过设定规则评判,挑选出训练好的大模型一组答案中的正面回答,以及需要训练的大模型一组答案中的负面回答,并且,正面回答要求具有步骤和逻辑,负面回答可以为问题分析不合理、归类错误、输出格式错误等问题,以此构建正负样本对并结合奖励模型继续对需要训练的大模型进行训练,以训练出满足需求的大模型部署在应用场景中使用即可。

[0067] 综上,通过设定多个思维链数据,且支持在模型训练阶段,按照问题类型选择,从而可以有效的保证对样本问题的更新不会篡改问题自身含义,并且可以达到强化学习的目的,从而提高模型训练精度。

[0068] 步骤S204,利用所述思维链数据将所述样本问题更新为目标样本问题,并将所述目标样本问题输入至初始问答模型进行处理,获得预测答案。

[0069] 具体的,在上述从问答样本对中读取到样本问题及其对应的思维链数据后,进一步的,为了能够使得问答模型可以学习到思考过程,并且提高奖励模型打分时可以针对不同长度问题给出精准分值,此时可以先利用思维链数据将样本问题更新为目标样本问题,实现在样本问题中添加促使模型可以学习的思考知识,此后即可将目标样本问题输入至初始问答模型进行处理,从而获得预测答案后,再利用奖励模型进行打分,从而结合打分结果完成模型优化处理。

[0070] 其中,目标样本问题具体是指利用思维链数据对样本问题进行更新后得到的问题文本,该目标样本问题中携带有使问答模型学习思考知识的文本内容;相应的,初始问答模型具体是指能够基于问题给出答案的大语言模型。相应的,预测答案具体是指初始问答模型针对目标样本问题输出的预测答案文本。需要说明的是,由于目标样本问题中携带有思维链数据对应的文本内容,且该文本内容的目的在于使得模型可以学习思考知识,因此,初

始问答模型输出的预测答案中,将具有样本问题对应的预测答案文本,同时还具有样本问题的解答思路文本,以便后续可以结合目标样本问题和当前得到的预测答案完成精度更高的打分,以使得问答模型可以从分值确定需要强化和削弱的参数,以保证模型预测精度和泛化能力。

[0071] 进一步的,在更新样本问题时,为了避免更新时出错,影响样本问题自身所需要表达的问题内容,可以采用字单元匹配的方式实现。本实施例中,具体实现方式如下:

[0072] 在所述样本问题中确定与所述思维链数据匹配的待变更字单元,利用所述思维链数据对所述待变更字单元进行更新,获得目标样本问题。

[0073] 具体的,待变更字单元具体是指样本问题中与思维链数据匹配且需要进行变更的字单元。基于此,在得到思维链数据后,可以在样本问题中确定与思维链数据匹配的待变更字单元,此时再利用思维链数据对待变更字单元进行更新,根据更新结果即可获得目标样本问题,此后再进行模型的训练即可。

[0074] 沿用上例,在得到样本问题{1*1+2*2的结果是多少?只回答结果}以及思维链数据{需要回答思考过程}后,此时可以利用思维链数据{需要回答思考过程}对样本问题{1*1+2*2的结果是多少?只回答结果}进行更新,根据更新结果,将得到目标样本问题{1*1+2*2的结果是多少?需要回答思考过程}。进一步的,可以将目标样本问题输入至问答模型进行处理,此时得到的预测答案为{首先我们计算第一步1*1=1,第二步计算2*2=4,第三步将前两步结果加到一起1+4=5,所以最终结果为5},得到预测答案中不仅包含问题答案,还包括问题的解题步骤,此后再利用奖励模型在此基础上进行打分,即可实现对问答模型的强化学习,提高模型回答精度。

[0075] 综上,通过匹配待变更字单元的方式进行样本问题更新,可以确保更新后的目标样本问题与原始样本问题的核心内容不变,但是可以具有使得模型学习的思考知识,从而达到不同长度的样本都可以提高模型预测精度的目的。

[0076] 步骤S206,利用所述初始问答模型关联的奖励模型根据所述问答样本对中的样本答案,对所述预测答案进行打分,获得优化分值。

[0077] 具体的,在上述得到初始问答模型输出的预测答案后,进一步的,考虑到奖励模型的目的是在强化阶段促使初始问答模型可以更准确的进行优化,因此奖励模型的打分结果是决定问答模型优化好坏的标准,而为了让能够提高奖励模型的打分精准度,在进行打分时,奖励模型将结合包含步骤内容和答案内容的预测答案完成打分。也就是说,可以利用初始问答模型关联的奖励模型根据问答样本对中的样本答案,对预测答案进行打分,从而输出能够反映模型当前预测能力的优化分值,以便后续可以结合优化分值完成模型的调参。其中,优化分值越高,表明模型预测精度越高,反之,优化分值越低,表明模型预测精度越低,并且,优化分值除采用整体性分值外,还可以是每个token对应分值组成的序列,从而能够更细粒度的表明初始问答模型在每个token上的预测精度,以便后续优化时更具有针对性。

[0078] 其中,奖励模型即为针对初始问答模型在强化学习阶段部署且训练好的奖励模型,并且该奖励模型的输入为预测答案、目标样本问题和样本答案的拼接结果,输出为优化分值。

[0079] 进一步的,在利用奖励模型进行打分时,为了能够使得奖励模型可以给出精度更

高且可解释性更强的优化分值,可以通过拼接问题和答案后输入奖励模型进行打分。本实施例中,具体实现方式如下:

[0080] 加载所述初始问答模型关联的奖励模型,以及将所述目标样本问题和所述预测答案拼接为优化文本;在所述问答样本对中提取样本答案,并利用所述奖励模型根据所述样本答案对所述优化文本进行打分,获得优化分值。

[0081] 具体的,优化文本具体是指对目标样本问题和预测答案进行拼接后得到的文本内容,用于输入奖励模型,利用奖励模型评估当前问答模型的输出结果准确性高低。

[0082] 基于此,在得到包含步骤信息和答案信息的预测答案后,此时可以先加载初始问答模型关联的奖励模型,同时可以将目标样本问题和预测答案进行拼接,根据拼接结果可以得到优化文本;再从问答样本对中提取样本答案;此后即可利用奖励模型根据样本答案对优化文本进行打分,即可优化分值,以便后续进行模型优化使用。

[0083] 沿用上例,在得到预测答案{首先我们计算第一步1*1=1,第二步计算2*2=4,第三步将前两步结果加到一起1+4=5,所以最终结果为5}后,可以将预测答案与目标样本问题{1*1+2*2的结果是多少?需要回答思考过程}进行拼接,得到优化文本{1*1+2*2的结果是多少?需要回答思考过程;首先我们计算第一步1*1=1,第二步计算2*2=4,第三步将前两步结果加到一起1+4=5,所以最终结果为5}。进一步的,再从问答样本对中提取样本答案{5}结合优化文本一同输入至奖励模型进行打分;假设问答模型针对问题给出的一组答案分别为5和3,此时两个预测答案中的每个token的分值分布如图3所示。通过token的分值可以体现问答模型的预测能力强弱,以便于后续进行精准的模型优化。

[0084] 综上,通过拼接为优化文本使用奖励模型进行打分,可以保证奖励模型针对预测答案给出更为准确的分值,该分值能够体现问答模型的预测能力强弱,因此后续可以更为精准的对模型进行调优。

[0085] 此外,为了可以使得奖励模型能够充分且精准反映出不同长度文本对应的分值,提高奖励模型泛化能力,从而促使问答模型可以得到更好的强化学习,可以通过如下方式训练奖励模型。本实施例中,具体实现方式如下:

[0086] 获取奖励样本以及所述奖励样本对应的分值向量样本序列,其中,所述分值向量样本序列包含所述奖励样本中每个字单元对应的标准分值向量;将所述奖励样本输入至初始奖励模型进行打分,获得分值向量预测序列,其中,所述分值向量预测序列包含所述奖励样本中每个字单元对应的预测分值向量;根据所述分值向量样本序列和所述分值向量预测序列计算损失值,并基于所述损失值对所述初始奖励模型进行调参,直至获得满足奖励训练停止条件的奖励模型。

[0087] 具体的,奖励样本具体是指由问题文本、答案文本以及真实答案文本组成的正/负样本,用于训练奖励模型使用。相应的,分值向量样本序列具体是指由奖励样本中的每个字单元对应的标准分值向量组成的序列,并且每个字单元对应的标准分值向量即为每个字单元对应的真实分值的向量表达。相应的,分值向量预测序列具体是指奖励模型针对每个字单元进行打分后,得到的预测分值向量组成的序列。也就是说,分值向量预测序列中包含的预测分值向量为每个字单元对应的预测分值的向量表达。

[0088] 基于此,在奖励模型的训练阶段,为了能够使得奖励模型可以具有更为精准的打分能力,从而促使问答模型可以在强化学习阶段得到更为准确的调优,可以先获取奖励样

本以及奖励样本对应的分值向量样本序列,其中,分值向量样本序列由奖励样本中每个字单元对应的标准分值向量组成;此后,可以将奖励样本输入至初始奖励模型进行打分,此时将得到奖励模型针对奖励样本进行打分后得到的分值向量预测序列,其中,分值向量预测序列由奖励样本中每个字单元对应的预测分值向量组成。在得到奖励模型输出的分值向量预测序列后,即可结合分值向量样本序列和分值向量预测序列进行损失值的计算,此后再基于损失值对初始奖励模型进行调参,直至获得满足奖励训练停止条件的奖励模型即可。

[0089] 实际应用中,在结合预测结果和标签计算损失值时,通过如下公式(1)的损失函数完成计算:

[0090] $$\mathrm{Loss} = -\mathrm{mean}\left(\log\left(\sigma\left(R_{chosen} - R_{rejected}\right)\right)\right) \quad (1)$$

[0091] 其中,Loss为损失值,R_{chosen}和$R_{rejected}$为奖励模型针对奖励样本中每个token进行打分后输出的预测分值向量以及真实分值向量(如果长度不同,则可以使用pad强制补全),σ(.)表示sigmoid函数。

[0092] 此外,奖励训练停止条件具体是指停止训练奖励模型的条件,其包括但不限于损失值比较条件、验证集验证条件或者迭代次数条件,具体实施时,可以根据实际需求选择,本实施例在此不做任何限定。

[0093] 综上,通过对奖励模型进行训练,可以使得奖励模型在进行预测答案进行打分时,具有更为精准的分值输出,从而使得问答模型在强化学习阶段根据奖励模型输出更为精准的分值完成调参。

[0094] 进一步的,考虑到奖励模型在训练阶段中使用的损失函数更为关注输入的最后一个token的分值,因此为了能够强化这一特性,并且可以支持奖励模型针对不同长度文本都可以给出准确度更高的分值,可以采用包含向量约束项的损失函数完成损失计算。本实施例中,具体实现方式如下:

[0095] 确定所述初始奖励模型预设的目标损失函数,其中,所述目标损失函数包含对目标分值向量进行约束的向量约束项;利用所述目标损失函数对所述分值向量样本序列和所述分值向量预测序列进行计算,获得损失值,并执行基于所述损失值对所述初始奖励模型进行调参的步骤。

[0096] 具体的,目标损失函数具体是指包含对目标分值向量进行约束的向量约束项的损失函数,该损失函数可以理解为在上述公式(1)的基础上添加向量约束项后得到的损失函数。其中,目标分值向量具体是指奖励模型输入的奖励样本中最后一个token对应的分值向量,向量约束项即为针对最后一个token对应的分值向量进行约束的计算项。

[0097] 基于此,结合上述公式(1)可知,该损失函数主要在于增加整体打分向量之间的距离,但是后续强化学习训练过程中,问答模型实际上是着重利用最后一个token的打分向量,因此为了能够增加对最后一个token 打分数值的差别的鼓励,可以在损失函数基础上增加向量约束项,此后即可利用目标损失函数对分值向量样本序列和分值向量预测序列进行计算,获得损失值,并执行基于损失值对初始奖励模型进行调参的步骤即可。

[0098] 实际应用中,在公式(1)损失函数的基础上增加对最后一个token的分值向量的差别的鼓励后,可以得到如公式(2)对应的损失函数:

[0099] $$\mathrm{Loss} = -\mathrm{mean}\left(\log\left(\sigma\left(R_{chosen} - R_{rejected}\right)\right)\right) - \lambda \times \log\left(\sigma\left(R_{chosen}[-1] - R_{rejected}[-1]\right)\right) \quad (2)$$

[0100] 其中,λ的取值可以根据实际需求设定,比如0.4等,本实施例在此不做任何限定。

[0101] 结合公式(1)和(2)可知,考虑到奖励模型理论状态下会对正样本每个token打高分,负样本每个token打低分;但是实际上由于奖励模型未经过充分训练,可能存在正样本的token平均分值高于负样本平均分值,但是单独个别token对应的分值可能存在正样本分值低于负样本分值的情况,因此为了能够约束这一情况,在公式(1)损失函数的基础上,增加了向量约束项,用于考虑在强化学习阶段着重应用最后一个token分值向量的前提下,增加对最后一个token分值向量的约束,从而确保公式(2)对应的损失函数可以有效的调参奖励模型,使其应用于问答模型训练阶段,可以具有更精准的打分能力。

[0102] 综上,通过采用包含向量约束项的目标损失函数对奖励模型进行调参,可以实现在损失函数中增加在强化学习阶段着重考虑的分值向量的约束项,以此为基础对奖励模型进行优化,可以使得奖励模型在进行打分时,能够结合该约束给出合理分值,从而提高强化学习阶段的模型优化效果。

[0103] 更进一步的,为了提高奖励模型的泛化能力,还可以结合SimCSE优化损失函数。本实施例中,具体实现方式如下:

[0104] 确定所述初始奖励模型预设的初始损失函数;响应于针对所述初始奖励模型提交的模型优化请求对所述初始损失函数进行更新,其中,所述模型优化请求用于在所述初始损失函数中添加语义学习信息;根据更新结果确定目标损失函数,并执行确定所述初始奖励模型预设的目标损失函数的步骤。

[0105] 具体的,优化请求具体是指在得到初始损失函数后,针对初始损失函数添加语义学习信息的请求,该请求用于在奖励模型优化阶段更改损失函数,使得更改后的损失函数可以使得奖励模型具有更好的泛化能力。

[0106] 基于此,可以先确定初始奖励模型预设的初始损失函数;此后可以响应于针对初始奖励模型提交的模型优化请求对初始损失函数进行更新,并且该模型优化请求用于在初始损失函数中添加语义学习信息;此后即可根据更新结果确定目标损失函数,并执行确定所述初始奖励模型预设的目标损失函数的步骤即可。

[0107] 实际应用中,在针对如上公式(2)对应的损失函数进行优化时,可以采用SimCSE(Simple Contrastive Learning of Sentence Embeddings,简单对比学习的句子嵌入)方法。其中,SimCSE的主要思想是通过对比学习(Contrastive Learning)来训练句子嵌入模型。具体来说,它会将同一个句子的两个独立副本作为正样本对,然后通过最大化这两个副本嵌入向量的余弦相似度,并最小化与其他句子(负样本)的相似度,来训练模型。SimCSE的优点在于其简单且有效。它不需要复杂的数据预处理或标签,只需要原始的文本数据即可。此外,SimCSE还可以与预训练语言模型(如BERT、RoBERTa等)结合使用,进一步提升性能。其中,SimCSE的损失函数是一个对比损失函数,具体形式如下公式(3):

[0108] $L_{simcse} = -\log(\exp(sim(u, v) / \tau) / \Sigma \exp(sim(u, v') / \tau))$ (3)

[0109] 其中,u和v是同一个句子的两个独立副本的嵌入向量,v'是其他句子的嵌入向量,sim(u, v)表示u和v的余弦相似度,τ是一个温度参数,Σ表示对所有负样本对的求和。该损失函数的目标是最大化正样本对的相似度,并最小化与负样本对的相似度。通过优化该损失函数,模型可以学会生成能够准确表示句子语义的嵌入向量。因此,在上述公式(3)损失函数的基础上,对公式(2)损失函数进行优化,即可得到如公式(4)所示的目标损失函数:

[0110] $$Loss = -mean\left(\log\left(\sigma\left(R_{chosen} - R_{rejected}\right)\right)\right) - \lambda \times \log\left(\sigma\left(R_{chosen}[-1] - R_{rejected}[-1]\right)\right) + L_{simcse} \quad (4)$$

[0111] 通过上述公式(4)所示的损失函数,可以提高奖励模型的泛化能力,从而实现在强化学习阶段,可以提高模型的预测精度。

[0112] 步骤S208,基于所述优化分值对所述初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型。

[0113] 具体的,在上述得到奖励模型输出的优化分值后,进一步的,由于优化分值可以表现出当前问答模型预测能力的强弱,因此基于优化分值可以确定问答模型在强化学习阶段的调参方向,以此为基础可以基于优化分值对初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型即可。

[0114] 其中,训练停止条件具体是指停止训练问答模型的条件,其包括但不限于损失值比较条件、验证集验证条件或者迭代次数条件,具体实施时,可以根据实际需求选择,本实施例在此不做任何限定。

[0115] 进一步的,在模型调参阶段,可以结合优化分值和样本答案构建模型优化策略,基于该策略可以明确问答模型所需要强化和削弱的参数,从而提高模型预测精度。本实施例中,具体实现方式如下:

[0116] 根据所述优化分值和所述样本答案针对所述初始问答模型构建模型优化策略;按照所述模型优化策略对所述初始问答模型进行调参,并检测调参后的初始问答模型是否满足训练停止条件;若否,将调参后的初始问答模型作为初始问答模型,并执行在问答样本对中提取样本问题的步骤;若是,将调参后的初始问答模型作为目标问答模型。

[0117] 具体的,模型优化策略具体是指结合优化分值以及样本答案对初始问答模型进行规划的优化策略,该策略可以强化模型中需要提升的参数,削弱模型中过强的参数,从而提高模型预测精度,且避免模型过拟合。

[0118] 基于此,在模型调参阶段,可以先根据优化分值和样本答案针对初始问答模型构建模型优化策略;此后可以按照模型优化策略对所述初始问答模型进行调参,并检测调参后的初始问答模型是否满足训练停止条件;若否,说明问答模型还需要继续进行训练,因此可以将调参后的初始问答模型作为初始问答模型,并返回执行步骤S202;直至某一调参处理后,模型满足训练停止条件,即可将调参后的初始问答模型作为目标问答模型,以部署在具体业务场景中使用即可。

[0119] 沿用上例,在得到奖励模型针对优化文本{1*1+2*2的结果是多少?需要回答思考过程;首先我们计算第一步1*1=1,第二步计算2*2=4,第三步将前两步结果加到一起1+4=5,所以最终结果为5}进行打分后得到的优化分值后,此时可以结合优化分值对问答模型进行调参,调参后通过验证集对问答模型进行验证,确定其还未达到期望的预测精度,则此时还需要选择新的样本继续进行训练,直至通过验证集进行验证后确定其满足期望预测精度时,即可将问答模型部署到下游业务场景,用于为用户提供问答服务。

[0120] 本实施例提供的问答模型训练方法,为了能够在问答模型训练过程中,可以通过奖励模型精准反映当前样本对训练问答模型后问答模型的预测精度,可以先在问答样本对中提取样本问题,并确定样本问题对应的思维链数据;在此基础上,为了能够使得简短问答依旧可以强化训练问答模型,此时可以利用思维链数据将样本问题更新为目标样本问题,实现通过目标样本问题体现更多的文本内容,此后即可将目标样本问题输入至初始问答模

型进行处理,获得预测答案;在此基础上,由于样本问题被更新为目标样本问题,因此奖励模型可以在进行模型预测精度打分时,可以给出能够精准反映模型预测精度的分值,即利用初始问答模型关联的奖励模型根据问答样本对中的样本答案,对预测答案进行打分,即可获得优化分值;最后再基于优化分值对初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型即可。实现在问答模型的强化学习阶段,可以通过思维链数据更新样本问题,使得样本问题携带更丰富的文本信息,以此经过问答模型预测后,可以使得奖励模型针对样本问题给出更加精准的打分,从而通过打分结果精准反映出问答模型需要优化的程度,以此为基础进行后续训练,可以有效的提高模型训练精度,以实现在应用阶段,可以满足下游业务使用。

[0121] 参见图4,图4示出了根据本说明书一个实施例提供的一种文本处理方法的流程图,具体包括以下步骤。

[0122] 步骤S402,接收客户端上传的问题文本;

[0123] 步骤S404,确定所述问题文本对应的问题领域信息,并选择所述问题领域信息匹配的目标问答模型,其中,所述目标问答模型通过所述问答模型训练方法获得;

[0124] 步骤S406,将所述问题文本输入至所述目标问答模型进行处理,获得答案文本,并将所述答案文本反馈至所述客户端。

[0125] 具体的,问题文本具体是指由客户端上传的待答复文本。相应的,问题领域信息具体是指描述问题所属领域的信息,且不同的领域可以部署不同的问答模型,从而使得问答模型可以专项进行问题答复。相应的,答案文本具体是指针对问题文本进行预测后得到的答案。

[0126] 基于此,在接收到客户端上传的问题文本后,可以先确定问题文本对应的问题领域信息,此时可以选择问题领域信息匹配的目标问答模型,而后,将问题文本输入至目标问答模型进行处理,即可获得答案文本,并将其反馈至客户端供用户参考即可。

[0127] 比如,用户输入问题文本为{A歌曲的作者是谁},此时可以选择匹配音乐领域的问答模型进行处理,获得答案文本{甲},同时,还可以增加关于用户甲的介绍,并将其反馈至用户客户端,供用户观看即可,用户在客户端观看到的答案内容可以是{A歌曲的作者是甲,男性,年龄**岁,出生于****地,代表作有****}。

[0128] 参见图5,图5示出了根据本说明书一个实施例提供的一种奖励模型训练方法的流程图,具体包括以下步骤。

[0129] 步骤S502,获取奖励样本以及所述奖励样本对应的分值向量样本序列,其中,所述分值向量样本序列包含所述奖励样本中每个字单元对应的标准分值向量;

[0130] 步骤S504,将所述奖励样本输入至初始奖励模型进行打分,获得分值向量预测序列,其中,所述分值向量预测序列包含所述奖励样本中每个字单元对应的预测分值向量;

[0131] 步骤S506,确定所述初始奖励模型预设的目标损失函数,其中,所述目标损失函数包含对目标分值向量进行约束的向量约束项;

[0132] 步骤S508,利用所述目标损失函数对所述分值向量样本序列和所述分值向量预测序列进行计算,根据计算结果对所述初始奖励模型进行调参,直至获得满足奖励训练停止条件的奖励模型。

[0133] 需要说明的是,本实施例提供的奖励模型训练方法为上述问答模型训练方法中应

用的奖励模型对应的训练方法,本实施例提供的奖励模型训练方法的描述可参见上述实施例中关于奖励模型训练的描述,本实施例在此不做任何过多赘述。

[0134] 下述结合附图6,以本说明书提供的问答模型训练方法在问答交互场景中的应用为例,对所述问答模型训练方法进行进一步说明。其中,图6示出了本说明书一个实施例提供的一种问答模型训练方法的处理过程流程图,具体包括以下步骤。

[0135] 步骤S602,在问答样本对中提取样本问题。

[0136] 步骤S604,确定样本问题对应的问题类型,在目标数据库中选择问题类型匹配的候选思维链数据,作为思维链数据。

[0137] 步骤S606,在样本问题中确定与思维链数据匹配的待变更字单元。

[0138] 步骤S608,利用思维链数据对待变更字单元进行更新,获得目标样本问题。

[0139] 步骤S610,将目标样本问题输入至初始问答模型进行处理,获得预测答案。

[0140] 步骤S612,加载初始问答模型关联的奖励模型,以及将目标样本问题和预测答案拼接为优化文本。

[0141] 步骤S614,在问答样本对中提取样本答案,并利用奖励模型根据样本答案对优化文本进行打分,获得优化分值。

[0142] 步骤S616,基于优化分值对初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型。

[0143] 步骤S618,接收客户端上传的问题文本。

[0144] 步骤S620,将问题文本输入至目标问答模型进行处理,获得答案文本,并将答案文本反馈至所述客户端。

[0145] 综上所述,为了能够在问答模型训练过程中,可以通过奖励模型精准反映当前样本对训练问答模型后问答模型的预测精度,可以先在问答样本对中提取样本问题,并确定样本问题对应的思维链数据;在此基础上,为了能够使得简短问答依旧可以强化训练问答模型,此时可以利用思维链数据将样本问题更新为目标样本问题,实现通过目标样本问题体现更多的文本内容,此后即可将目标样本问题输入至初始问答模型进行处理,获得预测答案;在此基础上,由于样本问题被更新为目标样本问题,因此奖励模型可以在进行模型预测精度打分时,可以给出能够精准反映模型预测精度的分值,即利用初始问答模型关联的奖励模型根据问答样本对中的样本答案,对预测答案进行打分,即可获得优化分值;最后再基于优化分值对初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型即可。实现在问答模型的强化学习阶段,可以通过思维链数据更新样本问题,使得样本问题携带更丰富的文本信息,以此经过问答模型预测后,可以使得奖励模型针对样本问题给出更加精准的打分,从而通过打分结果精准反映出问答模型需要优化的程度,以此为基础进行后续训练,可以有效的提高模型训练精度,以实现在应用阶段,可以满足下游业务使用。

[0146] 与上述方法实施例相对应,本说明书还提供了问答模型训练装置实施例,图7示出了本说明书一个实施例提供的一种问答模型训练装置的结构示意图。如图7所示,该装置包括:

[0147] 提取模块702,被配置为在问答样本对中提取样本问题,并确定所述样本问题对应的思维链数据;

[0148] 更新模块704,被配置为利用所述思维链数据将所述样本问题更新为目标样本问

题,并将所述目标样本问题输入至初始问答模型进行处理,获得预测答案;

[0149] 打分模块706,被配置为利用所述初始问答模型关联的奖励模型根据所述问答样本对中的样本答案,对所述预测答案进行打分,获得优化分值;

[0150] 调参模块708,被配置为基于所述优化分值对所述初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型。

[0151] 一个可选的实施例中,所述提取模块702进一步被配置为:

[0152] 确定所述样本问题对应的问题类型,在目标数据库中选择所述问题类型匹配的候选思维链数据,作为思维链数据;

[0153] 其中,所述更新模块704进一步被配置为:在所述样本问题中确定与所述思维链数据匹配的待变更字单元,利用所述思维链数据对所述待变更字单元进行更新,获得目标样本问题。

[0154] 一个可选的实施例中,所述打分模块706进一步被配置为:

[0155] 加载所述初始问答模型关联的奖励模型,以及将所述目标样本问题和所述预测答案拼接为优化文本;在所述问答样本对中提取样本答案,并利用所述奖励模型根据所述样本答案对所述优化文本进行打分,获得优化分值。

[0156] 一个可选的实施例中,所述调参模块708进一步被配置为:

[0157] 根据所述优化分值和所述样本答案针对所述初始问答模型构建模型优化策略;按照所述模型优化策略对所述初始问答模型进行调参,并检测调参后的初始问答模型是否满足训练停止条件;若否,将调参后的初始问答模型作为初始问答模型,并执行在问答样本对中提取样本问题的步骤;若是,将调参后的初始问答模型作为目标问答模型。

[0158] 一个可选的实施例中,所述装置还包括:

[0159] 奖励模型训练模块,被配置为获取奖励样本以及所述奖励样本对应的分值向量样本序列,其中,所述分值向量样本序列包含所述奖励样本中每个字单元对应的标准分值向量;将所述奖励样本输入至初始奖励模型进行打分,获得分值向量预测序列,其中,所述分值向量预测序列包含所述奖励样本中每个字单元对应的预测分值向量;根据所述分值向量样本序列和所述分值向量预测序列计算损失值,并基于所述损失值对所述初始奖励模型进行调参,直至获得满足奖励训练停止条件的奖励模型。

[0160] 一个可选的实施例中,所述奖励模型训练模块进一步被配置为:

[0161] 确定所述初始奖励模型预设的目标损失函数,其中,所述目标损失函数包含对目标分值向量进行约束的向量约束项;利用所述目标损失函数对所述分值向量样本序列和所述分值向量预测序列进行计算,获得损失值,并执行基于所述损失值对所述初始奖励模型进行调参的步骤。

[0162] 一个可选的实施例中,所述奖励模型训练模块进一步被配置为:

[0163] 确定所述初始奖励模型预设的初始损失函数;响应于针对所述初始奖励模型提交的模型优化请求对所述初始损失函数进行更新,其中,所述模型优化请求用于在所述初始损失函数中添加语义学习信息;根据更新结果确定目标损失函数,并执行确定所述初始奖励模型预设的目标损失函数的步骤。

[0164] 本实施例提供的问答模型训练装置,为了能够在问答模型训练过程中,可以通过奖励模型精准反映当前样本对训练问答模型后问答模型的预测精度,可以先在问答样本对

中提取样本问题,并确定样本问题对应的思维链数据;在此基础上,为了能够使得简短问答依旧可以强化训练问答模型,此时可以利用思维链数据将样本问题更新为目标样本问题,实现通过目标样本问题体现更多的文本内容,此后即可将目标样本问题输入至初始问答模型进行处理,获得预测答案;在此基础上,由于样本问题被更新为目标样本问题,因此奖励模型可以在进行模型预测精度打分时,可以给出能够精准反映模型预测精度的分值,即利用初始问答模型关联的奖励模型根据问答样本对中的样本答案,对预测答案进行打分,即可获得优化分值;最后再基于优化分值对初始问答模型进行调参,直至获得满足训练停止条件的目标问答模型即可。实现在问答模型的强化学习阶段,可以通过思维链数据更新样本问题,使得样本问题携带更丰富的文本信息,以此经过问答模型预测后,可以使得奖励模型针对样本问题给出更加精准的打分,从而通过打分结果精准反映出问答模型需要优化的程度,以此为基础进行后续训练,可以有效的提高模型训练精度,以实现在应用阶段,可以满足下游业务使用。

[0165] 上述为本实施例的一种问答模型训练装置的示意性方案。需要说明的是,该问答模型训练装置的技术方案与上述的问答模型训练方法的技术方案属于同一构思,问答模型训练装置的技术方案未详细描述的细节内容,均可以参见上述问答模型训练方法的技术方案的描述。

[0166] 与上述方法实施例相对应,本说明书还提供了文本处理装置实施例,图8示出了本说明书一个实施例提供的一种文本处理装置的结构示意图。如图8所示,该装置包括:

[0167] 接收模块802,被配置为接收客户端上传的问题文本;

[0168] 确定模块804,被配置为确定所述问题文本对应的问题领域信息,并选择所述问题领域信息匹配的目标问答模型,其中,所述目标问答模型通过所述问答模型训练方法获得;

[0169] 发送模块806,被配置为将所述问题文本输入至所述目标问答模型进行处理,获得答案文本,并将所述答案文本反馈至所述客户端。

[0170] 上述为本实施例的一种文本处理装置的示意性方案。需要说明的是,该文本处理装置的技术方案与上述的文本处理方法的技术方案属于同一构思,文本处理装置的技术方案未详细描述的细节内容,均可以参见上述文本处理方法的技术方案的描述。

[0171] 与上述方法实施例相对应,本说明书还提供了奖励模型训练装置实施例,图9示出了本说明书一个实施例提供的一种奖励模型训练装置的结构示意图。如图9所示,该装置包括:

[0172] 获取样本模块902,被配置为获取奖励样本以及所述奖励样本对应的分值向量样本序列,其中,所述分值向量样本序列包含所述奖励样本中每个字单元对应的标准分值向量;

[0173] 打分样本模块904,被配置为将所述奖励样本输入至初始奖励模型进行打分,获得分值向量预测序列,其中,所述分值向量预测序列包含所述奖励样本中每个字单元对应的预测分值向量;

[0174] 确定函数模块906,被配置为确定所述初始奖励模型预设的目标损失函数,其中,所述目标损失函数包含对目标分值向量进行约束的向量约束项;

[0175] 调参模型模块908,被配置为利用所述目标损失函数对所述分值向量样本序列和所述分值向量预测序列进行计算,根据计算结果对所述初始奖励模型进行调参,直至获得

满足奖励训练停止条件的奖励模型。

[0176] 上述为本实施例的一种奖励模型训练装置的示意性方案。需要说明的是,该奖励模型训练装置的技术方案与上述的奖励模型训练方法的技术方案属于同一构思,奖励模型训练装置的技术方案未详细描述的细节内容,均可以参见上述奖励模型训练方法的技术方案的描述。

[0177] 图10示出了根据本说明书一个实施例提供的一种计算设备1000的结构框图。该计算设备1000的部件包括但不限于存储器1010和处理器1020。处理器1020与存储器1010通过总线1030相连接,数据库1050用于保存数据。

[0178] 计算设备1000还包括接入设备1040,接入设备1040使得计算设备1000能够经由一个或多个网络1060通信。这些网络的示例包括公用交换电话网(PSTN,Public Switched Telephone Network)、局域网(LAN,Local Area Network)、广域网(WAN,Wide Area Network)、个域网(PAN,Personal Area Network)或诸如因特网的通信网络的组合。接入设备1040可以包括有线或无线的任何类型的网络接口(例如,网络接口卡(NIC,network interface controller))中的一个或多个,诸如IEEE802.11无线局域网(WLAN,Wireless Local Area Network)无线接口、全球微波互联接入(Wi-MAX,Worldwide Interoperability for Microwave Access)接口、以太网接口、通用串行总线(USB,Universal Serial Bus)接口、蜂窝网络接口、蓝牙接口、近场通信(NFC,Near Field Communication)。

[0179] 在本说明书的一个实施例中,计算设备1000的上述部件以及图10中未示出的其他部件也可以彼此相连接,例如通过总线。应当理解,图10所示的计算设备结构框图仅仅是出于示例的目的,而不是对本说明书范围的限制。本领域技术人员可以根据需要,增添或替换其他部件。

[0180] 计算设备1000可以是任何类型的静止或移动计算设备,包括移动计算机或移动计算设备(例如,平板计算机、个人数字助理、膝上型计算机、笔记本计算机、上网本等)、移动电话(例如,智能手机)、可佩戴的计算设备(例如,智能手表、智能眼镜等)或其他类型的移动设备,或者诸如台式计算机或个人计算机(PC,Personal Computer)的静止计算设备。计算设备1000还可以是移动式或静止式的服务器。

[0181] 其中,处理器1020用于执行如下计算机可执行指令,该计算机可执行指令被处理器执行时实现上述问答模型训练方法、文本处理方法或奖励模型训练方法的步骤。

[0182] 上述为本实施例的一种计算设备的示意性方案。需要说明的是,该计算设备的技术方案与上述的问答模型训练方法、文本处理方法或奖励模型训练方法的技术方案属于同一构思,计算设备的技术方案未详细描述的细节内容,均可以参见上述问答模型训练方法、文本处理方法或奖励模型训练方法的技术方案的描述。

[0183] 本说明书一实施例还提供一种计算机可读存储介质,其存储有计算机可执行指令,该计算机可执行指令被处理器执行时实现上述问答模型训练方法、文本处理方法或奖励模型训练方法的步骤。

[0184] 上述为本实施例的一种计算机可读存储介质的示意性方案。需要说明的是,该存储介质的技术方案与上述的问答模型训练方法、文本处理方法或奖励模型训练方法的技术方案属于同一构思,存储介质的技术方案未详细描述的细节内容,均可以参见上述问答模

型训练方法、文本处理方法或奖励模型训练方法的技术方案的描述。

[0185] 本说明书一实施例还提供一种计算机程序产品,包括计算机程序或指令,该计算机程序或指令被处理器执行时实现上述问答模型训练方法、文本处理方法或奖励模型训练方法的步骤。

[0186] 上述为本实施例的一种计算机程序产品的示意性方案。需要说明的是,该计算机程序产品的技术方案与上述的问答模型训练方法、文本处理方法或奖励模型训练方法的技术方案属于同一构思,计算机程序产品的技术方案未详细描述的细节内容,均可以参见上述问答模型训练方法、文本处理方法或奖励模型训练方法的技术方案的描述。

[0187] 上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下,在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外,在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中,多任务处理和并行处理也是可以的或者可能是有利的。

[0188] 所述计算机指令包括计算机程序代码,所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括:能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、电载波信号、电信信号以及软件分发介质等。需要说明的是,所述计算机可读介质包含的内容可以根据专利实践的要求进行适当的增减,例如在某些地区,根据专利实践,计算机可读介质不包括电载波信号和电信信号。

[0189] 需要说明的是,对于前述的各方法实施例,为了简便描述,故将其都表述为一系列的动作组合,但是本领域技术人员应该知悉,本说明书实施例并不受所描述的动作顺序的限制,因为依据本说明书实施例,某些步骤可以采用其它顺序或者同时进行。其次,本领域技术人员也应该知悉,说明书中所描述的实施例均属于优选实施例,所涉及的动作和模块并不一定都是本说明书实施例所必须的。

[0190] 在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中没有详述的部分,可以参见其它实施例的相关描述。

[0191] 以上公开的本说明书优选实施例只是用于帮助阐述本说明书。可选实施例并没有详尽叙述所有的细节,也不限制该发明仅为所述的具体实施方式。显然,根据本说明书实施例的内容,可作很多的修改和变化。本说明书选取并具体描述这些实施例,是为了更好地解释本说明书实施例的原理和实际应用,从而使所属技术领域技术人员能很好地理解和利用本说明书。

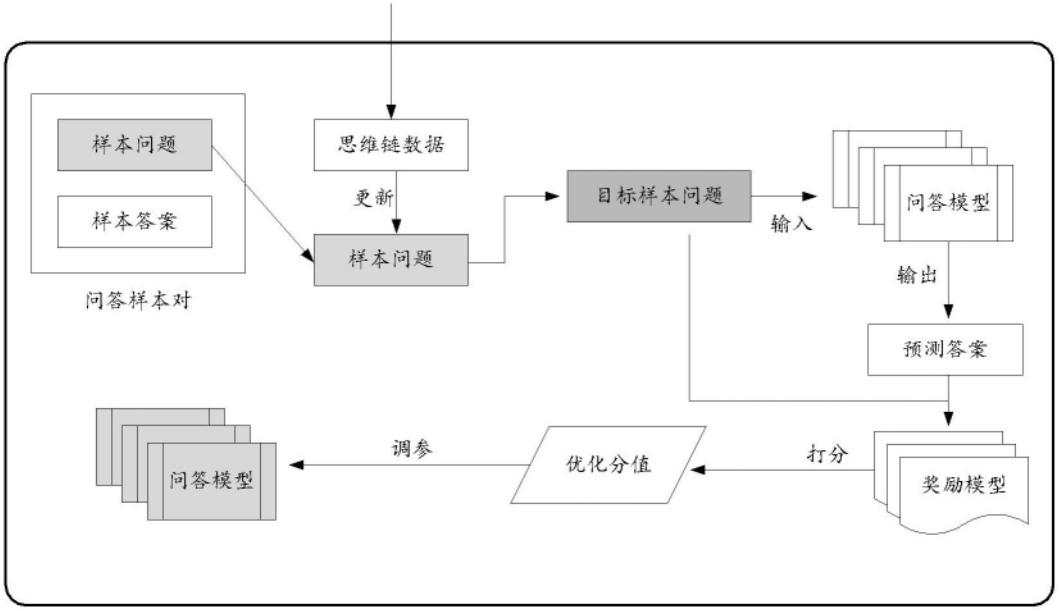

样本问题
样本答案
问答样本对
思维链数据
更新
样本问题
目标样本问题
输入
问答模型
输出
预测答案
打分
奖励模型
优化分值
调参
问答模型

图1

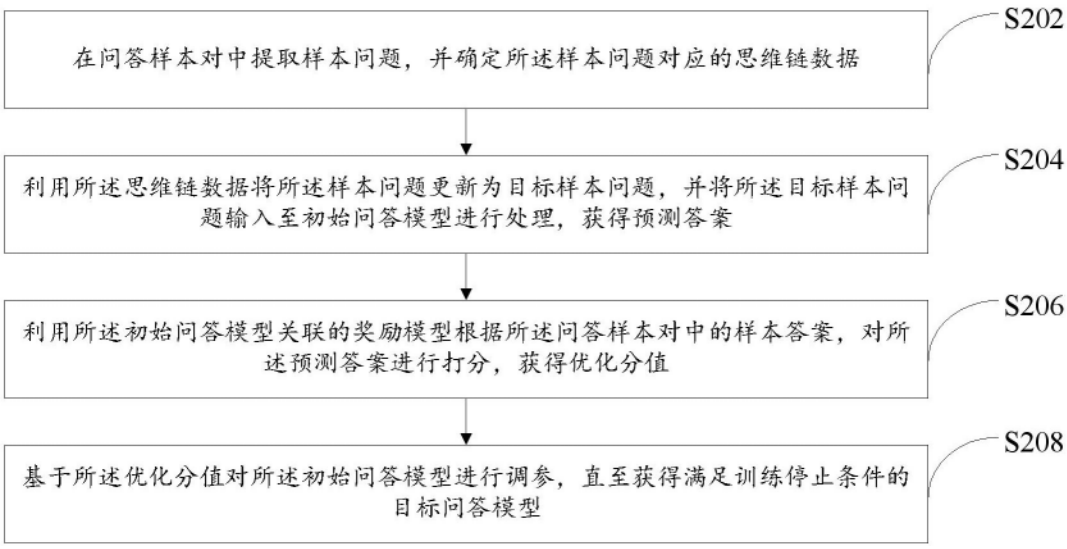

在问答样本对中提取样本问题，并确定所述样本问题对应的思维链数据
S202
利用所述思维链数据将所述样本问题更新为目标样本问题，并将所述目标样本问题输入至初始问答模型进行处理，获得预测答案
S204
利用所述初始问答模型关联的奖励模型根据所述问答样本对中的样本答案，对所述预测答案进行打分，获得优化分值
S206
基于所述优化分值对所述初始问答模型进行调参，直至获得满足训练停止条件的目标问答模型
S208

图2

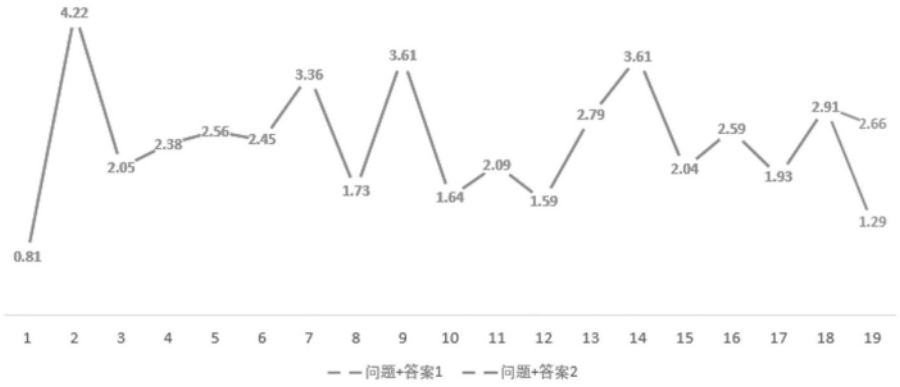

4.22
0.81
2.05
2.38
2.56
2.45
3.36
1.73
3.61
1.64
2.09
1.59
2.79
3.61
2.04
2.59
1.93
2.91
2.66
1.29
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
问题+答案1
问题+答案2

图3

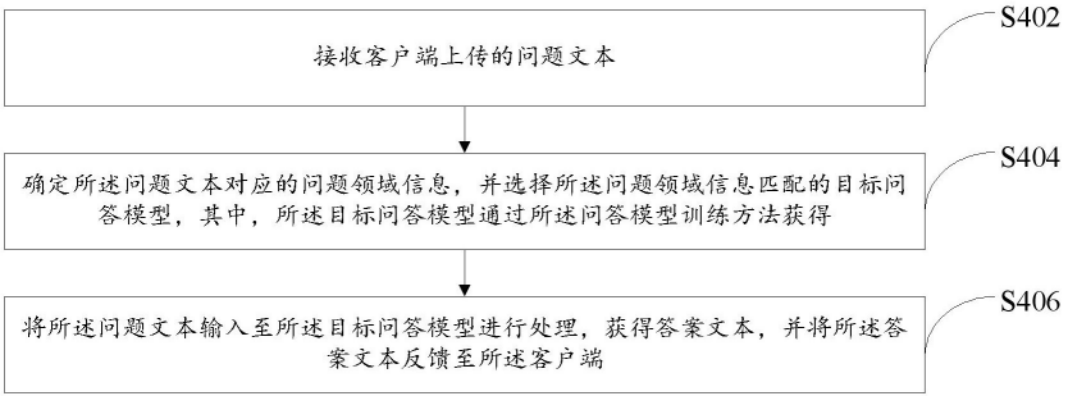

接收客户端上传的问题文本
S402
确定所述问题文本对应的问题领域信息，并选择所述问题领域信息匹配的目标问答模型，其中，所述目标问答模型通过所述问答模型训练方法获得
S404
将所述问题文本输入至所述目标问答模型进行处理，获得答案文本，并将所述答案文本反馈至所述客户端
S406

图4

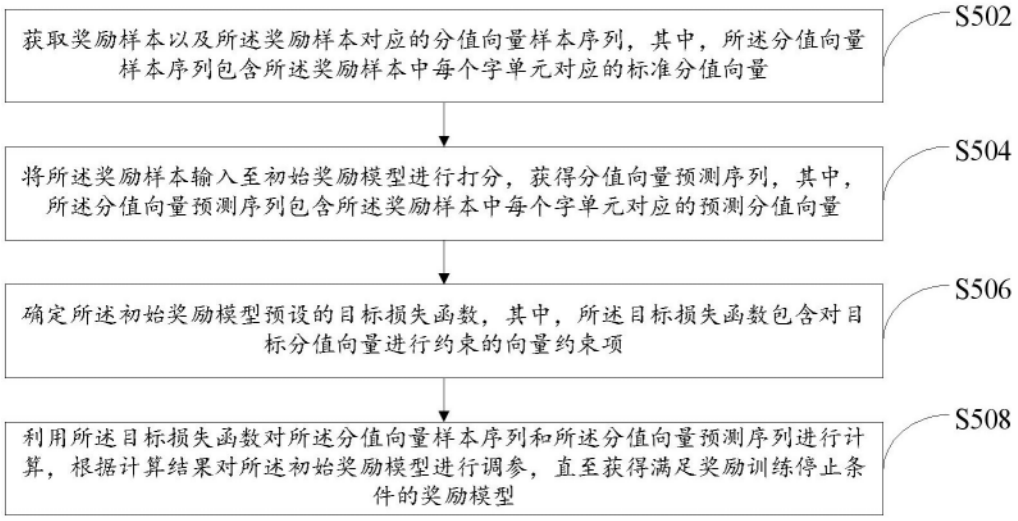

获取奖励样本以及所述奖励样本对应的分值向量样本序列，其中，所述分值向量样本序列包含所述奖励样本中每个字单元对应的标准分值向量
S502
将所述奖励样本输入至初始奖励模型进行打分，获得分值向量预测序列，其中，所述分值向量预测序列包含所述奖励样本中每个字单元对应的预测分值向量
S504
确定所述初始奖励模型预设的目标损失函数，其中，所述目标损失函数包含对目标分值向量进行约束的向量约束项
S506
利用所述目标损失函数对所述分值向量样本序列和所述分值向量预测序列进行计算，根据计算结果对所述初始奖励模型进行调参，直至获得满足奖励训练停止条件的奖励模型
S508

图5

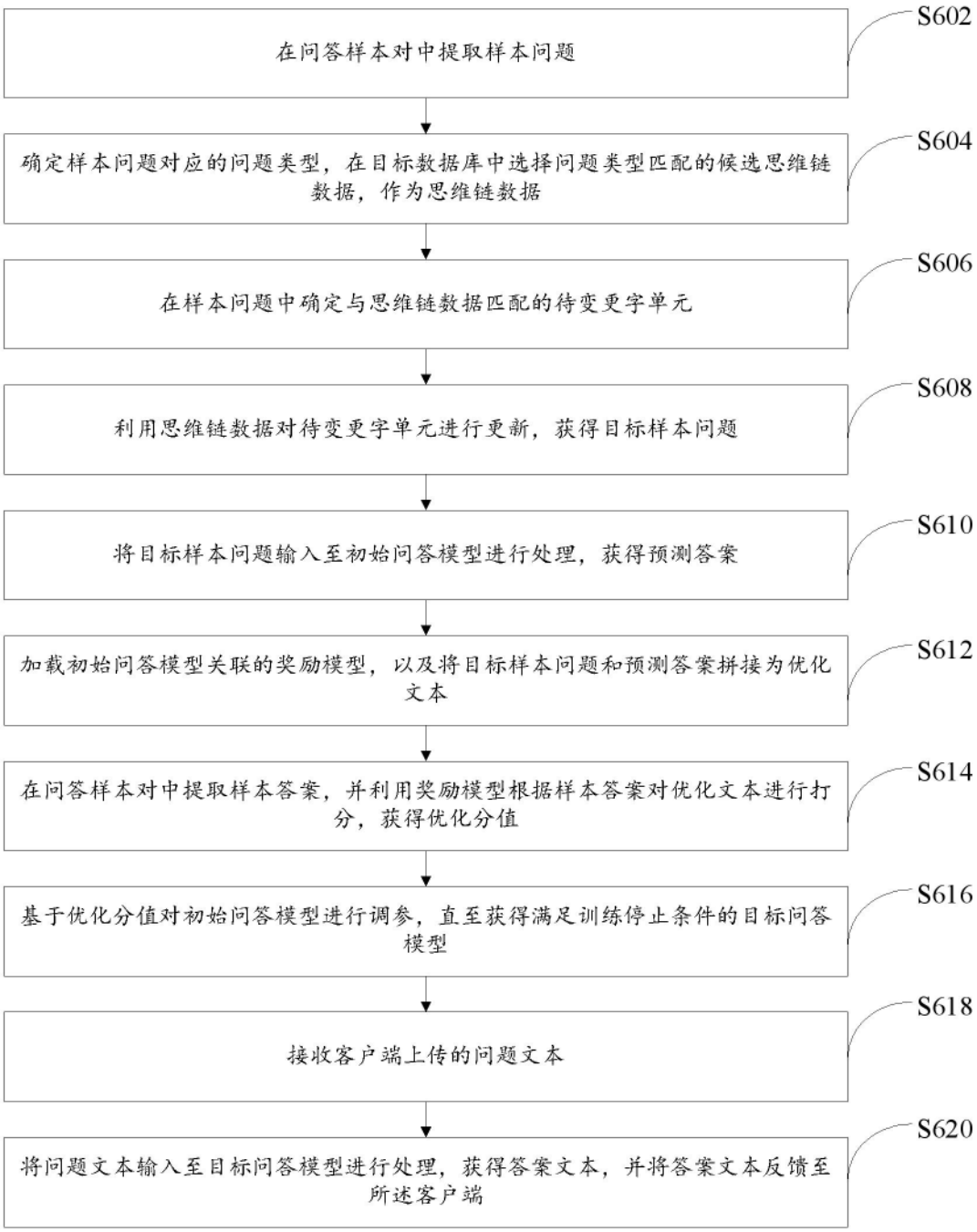

在问答样本对中提取样本问题
S602
确定样本问题对应的问题类型，在目标数据库中选择问题类型匹配的候选思维链数据，作为思维链数据
S604
在样本问题中确定与思维链数据匹配的待变更字单元
S606
利用思维链数据对待变更字单元进行更新，获得目标样本问题
S608
将目标样本问题输入至初始问答模型进行处理，获得预测答案
S610
加载初始问答模型关联的奖励模型，以及将目标样本问题和预测答案拼接为优化文本
S612
在问答样本对中提取样本答案，并利用奖励模型根据样本答案对优化文本进行打分，获得优化分值
S614
基于优化分值对初始问答模型进行调参，直至获得满足训练停止条件的目标问答模型
S616
接收客户端上传的问题文本
S618
将问题文本输入至目标问答模型进行处理，获得答案文本，并将答案文本反馈至所述客户端
S620

图6

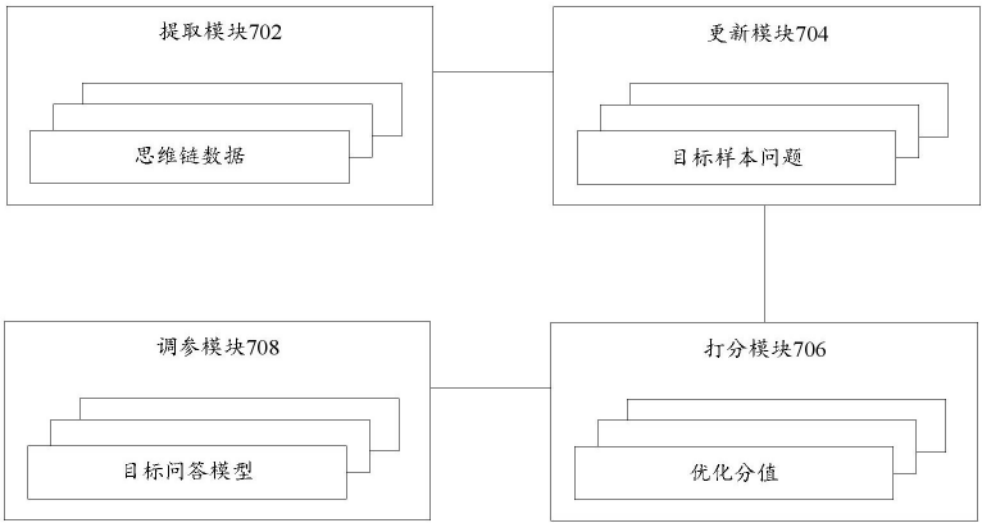

提取模块702
思维链数据
更新模块704
目标样本问题
调参模块708
目标问答模型
打分模块706
优化分值

图7

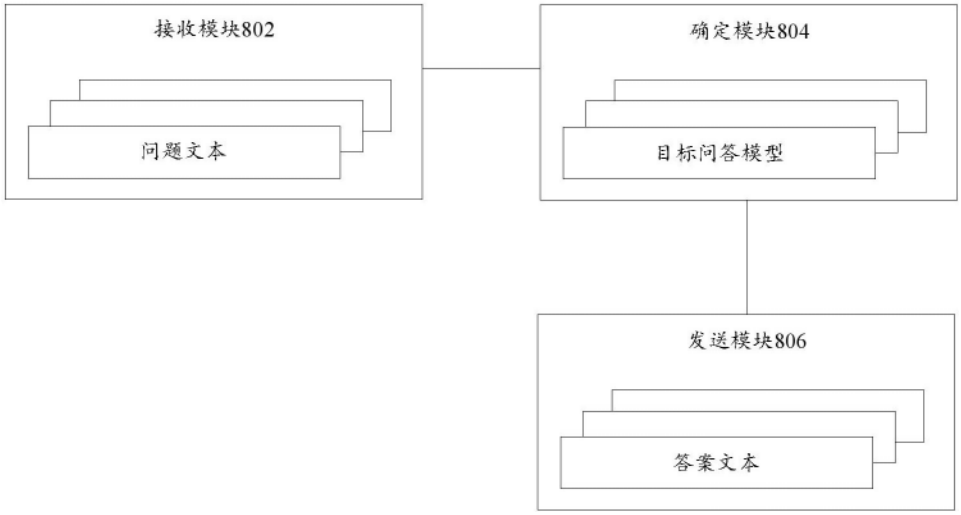

接收模块802
问题文本
确定模块804
目标问答模型
发送模块806
答案文本

图8

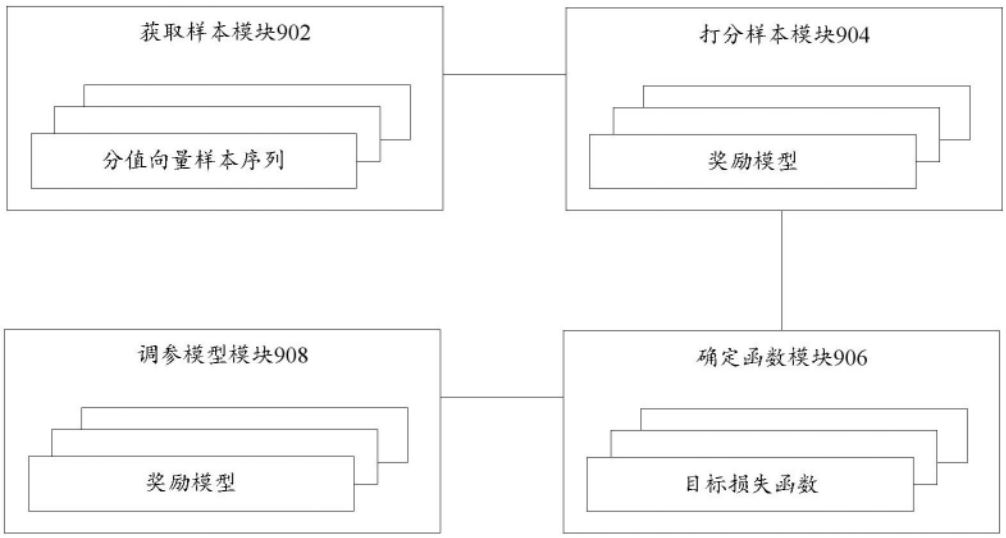

获取样本模块902
分值向量样本序列
打分样本模块904
奖励模型
调参模型模块908
奖励模型
确定函数模块906
目标损失函数

图9

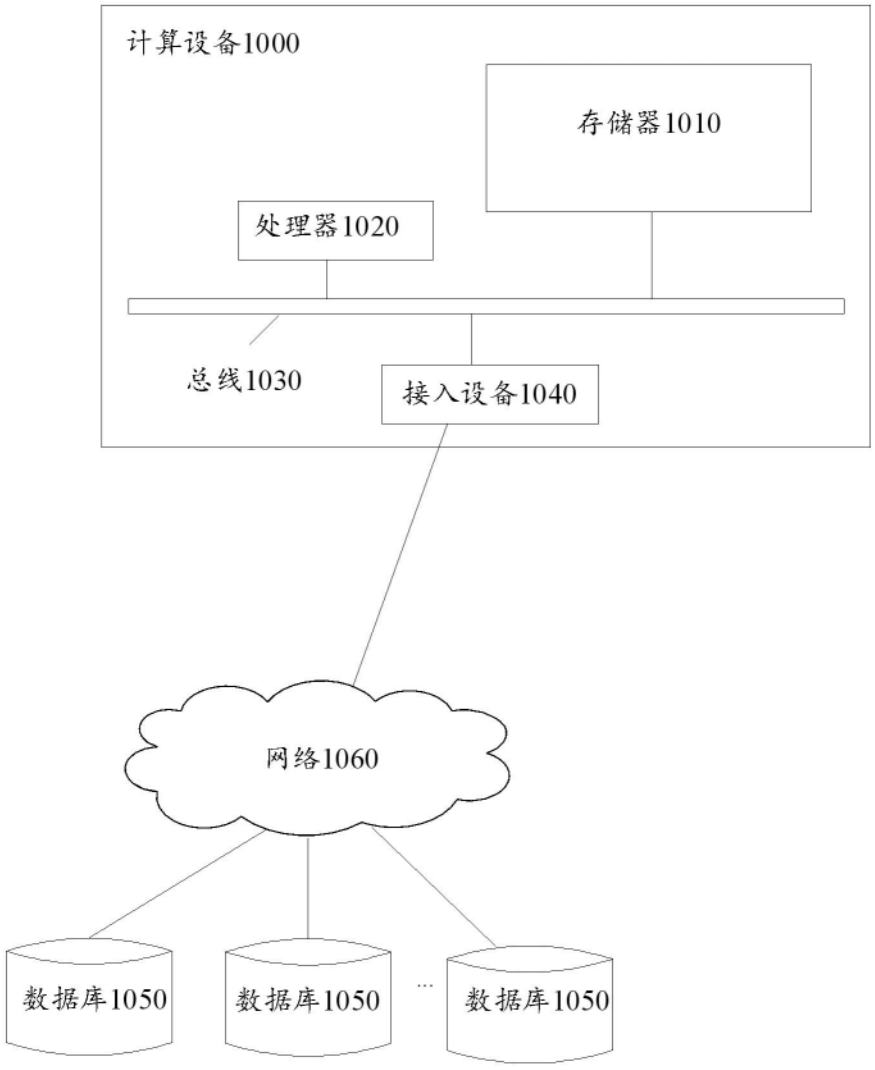

计算设备1000
存储器1010
处理器1020
总线1030
接入设备1040
网络1060
数据库1050
数据库1050
...
数据库1050

图10